A new standard for the analysis and design of replication studies

Held, Leonhard

Abstract: A new standard is proposed for the evidential assessment of replication studies. The approach combines a specific reverse Bayes technique with prior-predictive tail probabilities to define replication success. The method gives rise to a quantitative measure for replication success, called the sceptical p-value. The sceptical p-value integrates traditional significance of both the original and the replication study with a comparison of the respective effect sizes. It incorporates the uncertainty of both the original and the replication effect estimates and reduces to the ordinary p-value of the replication study if the uncertainty of the original effect estimate is ignored. The framework proposed can also be used to determine the power or the required replication sample size to achieve replication success. Numerical calculations highlight the difficulty of achieving replication success if the evidence from the original study is only suggestive. An application to data from the Open Science Collaboration project on the replicability of psychological science illustrates the methodology proposed.

DOI: https://doi.org/10.1111/rssa.12493

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 2, *pp.* 431–448

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting on 'Signs and sizes: understanding and replicating statistical findings' at the Society's 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President,* Professor D. Ashby, *in the Chair*]

Summary. A new standard is proposed for the evidential assessment of replication studies. The approach combines a specific reverse Bayes technique with prior-predictive tail probabilities to define replication success. The method gives rise to a quantitative measure for replication success, called the sceptical *p*-value. The sceptical *p*-value integrates traditional significance of both the original and the replication study with a comparison of the respective effect sizes. It incorporates the uncertainty of both the original and the replication effect estimates and reduces to the ordinary *p*-value of the replication study if the uncertainty of the original effect estimate is ignored. The framework proposed can also be used to determine the power or the required replication sample size to achieve replication success. Numerical calculations highlight the difficulty of achieving replication success if the evidence from the original study is only suggestive. An application to data from the Open Science Collaboration project on the replicability of psychological science illustrates the methodology proposed.

Keywords: Power; Prior–data conflict; Replication success; Reverse Bayes technique; Sample size; Sceptical *p*-value

1. Introduction

Replicability of research findings is crucial to the credibility of all empirical domains of science. As a consequence of the so-called replication crisis (Ioannidis, 2005; Begley and Ioannidis, 2015), recent years have witnessed increasing interest in large-scale replication projects, e.g. Open Science Collaboration (2015) and Camerer *et al.* (2016, 2018). Such efforts help to assess to what extent claims of new discoveries can be confirmed in independent replication studies whose procedures are as closely matched to the original studies as possible.

However, there is no established standard for the statistical evaluation of replication success. Standard significance of the replication study is often used as a criterion, but significance alone does not take the effect sizes of the original and replication study into account and can easily lead to conclusions that are the opposite to what the evidence warrants (Simonsohn, 2015). A comparison of the effect sizes of the original and replication study is also common, where a smaller replication effect estimate decreases the credibility of the original study result. A modification of this is to investigate whether the replication effect estimate is compatible with the original effect estimate (Bayarri and Mayoral, 2002; Patil *et al.*, 2016). Meta-analytic approaches take the results from the original and replication study at face value and combine them into an

Address for correspondence: Leonhard Held, Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland.
E-mail: leonhard.held@uzh.ch

overall effect size estimate. However, when conducting a replication, researchers are challenging the original study and asking *whether* they should take it at face value. This is an inherently asymmetric question, where exchangeability assumptions are not appropriate and alternative methods for evidence quantification are needed.

Recently the lack of a single accepted definition of replicability has been emphasized by Goodman *et al.* (2016) who called for a better understanding of the relationship between reproducibility and the truth of scientific claims. Researchers have started to develop Bayesian methods to analyse and design replication studies (Verhagen and Wagenmakers, 2014; van Aert and van Assen, 2017; Schönbrodt and Wagenmakers, 2018), but there is a lack of appropriate methodology based on traditional metrics (effect estimates, confidence intervals and p-values). To address this deficiency, I propose a principled approach, combining the analysis of credibility (Matthews, 2001a, b) with the prior criticism approach by Box (1980) and Evans and Moshonov (2006) to define *replication success* (Section 2). This gives rise to a new quantitative measure of replication success: the *sceptical p-value* (Section 3).

The sceptical p-value has attractive properties. It takes into account the results from both the original and the replication study and is always larger than the ordinary p-values from these studies. If the uncertainty of the original effect estimate is ignored, the sceptical p-value reduces to the ordinary p-value from the replication study. Moreover, the sceptical p-value considers replication studies with relatively small effect estimates (compared with the original estimates) as less successful. To avoid the so-called replication paradox (Ly *et al.*, 2018), a one-sided sceptical p-value is derived within the framework proposed, ensuring that replication success can occur only if the original and replication effect estimates have the same sign.

Statistical power is of central importance in assessing the reliability of science (Button *et al.*, 2013). Appropriate design of a replication study is key to tackling the replication crisis as many such studies are currently severely underpowered, even by traditional standards (Anderson and Maxwell, 2017). Methods to calculate the power for replication success are proposed in Section 4. Numerical calculations highlight the difficulty of achieving replication success if the evidence from the original study is only suggestive. The framework is also used to determine the sample size required to achieve replication success with appropriate power. Section 5 presents a reanalysis of data from the Open Science Collaboration (2015) project on the replicability of psychological science to illustrate the usefulness of the methodology proposed. I close with some comments in Section 6.

2. Assessment of replication success

Analysis of credibility (Matthews, 2001a, b) is a reverse Bayes procedure which was originally designed to assess the credibility of significant findings in the light of existing evidence. The discussion of Matthews (2001b) and his response provide additional insights on the philosophy and detail of this method. The idea to use Bayes's theorem in reverse originates in the work of I. J. Good (Good, 1950, 1983) and is increasingly used to assess the plausibility of scientific findings (Greenland, 2006, 2011; Held, 2013; Colquhoun, 2017, 2019).

2.1. Reverse Bayes analysis

Analysis of credibility combines a significant effect estimate from the original study with a sceptical prior (Spiegelhalter *et al.* (1994), section 4.1.3): a normal distribution centred near zero to represent doubts about large effect estimates. A sceptical prior shrinks the original effect estimate towards 0, where the amount of shrinkage depends on the sceptical prior variance.

Fletcher *et al.* (1993) have argued for the use of sceptical priors for original clinical study results, which often show a tendency for overoptimism.

To challenge the original study it is natural to ask how sceptical we would have to be not to find its apparently positive effect estimate convincing. This leads to a reverse Bayes approach, where the posterior is fixed to have a lower (or upper) credible limit exactly equal to 0 and the sceptical prior variance is chosen accordingly. The approach thus represents the objection by a sceptic who argues that the original result would no longer be 'significant' if combined with a sufficiently sceptical prior. The goal is now to persuade the sceptic by showing that this prior is unrealistic. To do so, a replication study is conducted. If the data from the replication study are in conflict with the sufficiently sceptical prior, the original study result is confirmed.

Suppose that the original study gives rise to a conventional confidence interval (CI) for the unknown effect size θ at level $1-\alpha$ with lower limit L and upper limit U. Assume that L and U are symmetric around the original point estimate $\hat{\theta}_o$ (assumed to be normally distributed) and that both are either positive or negative, i.e. the original effect is significant at significance level α. After a suitable transformation this framework covers a large number of commonly used effect measures such as differences in means, odds ratios, relative risks and correlations.

We first need to compute the variance of the sufficiently sceptical prior. Matthews (2001a) has shown that the equitailed credible interval of the sufficiently sceptical prior at level $1-\alpha$ has limits $\pm S$ where

$$S = \frac{(U-L)^2}{4\sqrt{(UL)}} \tag{1}$$

is the *scepticism limit* (Matthews, 2018). Note that equation (1) holds for any value of α, not just for the traditional 5% level. The sufficiently sceptical prior variance τ^2 can be derived from equation (1) and expressed as a function of the variance σ_o^2 (the squared standard error, which is assumed to be known) of the estimate $\hat{\theta}_o$, the corresponding test statistic $t_o = \hat{\theta}_o/\sigma_o$ and $z_{\alpha/2}$, the $(1-\alpha/2)$-quantile of the standard normal distribution (Held (2019), appendix):

$$\tau^2 = \frac{\sigma_o^2}{t_o^2/z_{\alpha/2}^2 - 1}, \tag{2}$$

where $t_o^2 > z_{\alpha/2}^2$ holds because of significance of the original study at level α.

Equation (2) shows that the sufficiently sceptical prior variance τ^2 can be both smaller or larger than σ_o^2, depending on the value of t_o^2. For a 'borderline' significant result where t_o^2 is close to $z_{\alpha/2}^2$, the sufficiently sceptical prior variance will be relatively large. If t_o^2 is substantially larger than $z_{\alpha/2}^2$, then the sufficiently sceptical prior variance will be relatively small.

The left-hand part of Fig. 1 shows an example of this procedure. The original study has effect estimate $\hat{\theta}_o = 0.57$ (95% CI from $L = 0.25$ to $U = 0.89$) and two-sided p-value $p_o = 0.0005$. The scepticism limit, calculated from equation (1), turns out to be $S = 0.22$.

2.2. Assessing prior–data conflict

The replication study that is shown in the right-hand part of Fig. 1 has an effect estimate of $\hat{\theta}_r = 0.33$ (95% CI from 0.01 to 0.65; $p_r = 0.046$). If the replication result is in conflict with the sufficiently sceptical prior, the original result is deemed credible. A visual comparison of the sufficiently sceptical prior with the replication study result in the right-hand part of Fig. 1 can be useful to assess potential conflict, but in general a more principled statistical approach is needed.

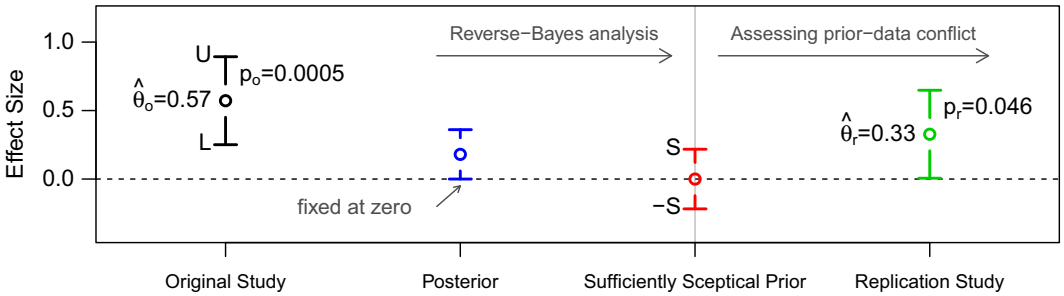

Fig. 1. Example of the assessment of replication success: the original study has effect estimate $\hat{\theta}_o = 0.57$ (95% CI from $L = 0.25$ to $U = 0.89$) and two-sided p-value $p_o = 0.0005$; the left-hand part of the figure illustrates the reverse Bayes derivation of the sufficiently sceptical prior with scepticism limit $S = 0.22$ based on the original study result and the posterior with lower credible limit fixed at zero; the comparison of the sufficiently sceptical prior with the replication study result ($\hat{\theta}_r = 0.33$; 95% CI from 0.01 to 0.65; $p_r = 0.046$) in the right-hand part of the figure is used to assess potential prior–data conflict

One option is to consider the original study as credible, if the absolute value of the effect estimate $\hat{\theta}_r$ from the replication study is larger than the scepticism limit S (Matthews, 2001a, b). In the above example the effect estimate in the replication study ($\hat{\theta}_r = 0.33$) is larger than the scepticism limit ($S = 0.22$), so the original study would be considered credible at the 5% level. However, a disadvantage of this approach is that it does not take the (known) variance σ_r^2 of the replication estimate $\hat{\theta}_r$ (in what follows also assumed to be normally distributed) into account. To address this issue, I propose to quantify prior–data conflict based on the prior-predictive distribution of $\hat{\theta}_r$: a normal distribution with mean 0 and variance $\tau^2 + \sigma_r^2$ (Spiegelhalter *et al.* (2004), section 5.8). This is the established way to check the compatibility of prior and data (Box, 1980; Evans and Moshonov, 2006) and leads to the test statistic

$$t_{\text{Box}} = \frac{\hat{\theta}_r}{\sqrt{(\tau^2 + \sigma_r^2)}} \tag{3}$$

and the tail probability $p_{\text{Box}} = \Pr\{\chi^2(1) \geqslant t_{\text{Box}}^2\}$ as the corresponding upper tail of a χ^2-distribution with 1 degree of freedom. Small values of p_{Box} indicate a conflict between the sufficiently sceptical prior and the estimate from the replication study and I define *replication success* at level α if $p_{\text{Box}} \leqslant \alpha$, or equivalently $t_{\text{Box}}^2 \geqslant z_{\alpha/2}^2$.

In the example that is shown in Fig. 1, the prior-predictive assessment of conflict gives $t_{\text{Box}} = 1.65$ with Box's tail probability $p_{\text{Box}} = 0.098 > 0.05$, so the replication study is not successful at the 5% level, although both the original and the replication study are significant at that level. This illustrates that replication success is a more stringent criterion than significance alone. For $\alpha = 10\%$, Box's tail probability is somewhat smaller ($p_{\text{Box}} = 0.078$), and we can declare replication success at the 10% level.

The example illustrates how Box's tail probability can be used to assess replication success at level α. However, it is difficult to interpret the actual value of p_{Box} as it depends on the choice of α. Furthermore, assessment of replication success is only possible if the original study result is significant at level α as otherwise the sufficiently sceptical prior would not exist and p_{Box} could not be computed. These issues motivate the work that is described in the next section where I introduce the sceptical p-value: a quantitative measure for replication success that is independent of the level α.

3. The sceptical *p*-value

Instead of dichotomizing replication studies into successful yes or no at some arbitrary level α, I now propose the *sceptical p-value* p_S to assess replication success quantitatively. The idea is to determine the largest confidence level $1 - p_S$ for the original confidence interval, at which we can declare replication success at level p_S. This parallels the duality of ordinary p-values and CIs, where the largest confidence level $1 - p$ at which we can declare significance can be used to compute the ordinary p-value p. Replication success at any prespecified level α is then equivalent to $p_S \leqslant \alpha$, just as significance at level α is equivalent to $p \leqslant \alpha$.

To determine p_S, let $c = \sigma_o^2/\sigma_r^2$ denote the ratio of the variances of the original and replication effect estimates and let $t_r = \hat{\theta}_r/\sigma_r$ denote the test statistic of the replication study. With equation (2) we can derive the prior-predictive variance of $\hat{\theta}_r$:

$$\tau^2 + \sigma_r^2 = \sigma_r^2 \left(\frac{c}{t_o^2/z_{\alpha/2}^2 - 1} + 1 \right). \tag{4}$$

Using equations (3) and (4), the requirement $t_{Box}^2 = \hat{\theta}_r^2/(\tau^2 + \sigma_r^2) \geqslant z_{\alpha/2}^2$ for replication success at level α can be written as

$$(t_o^2/z_{\alpha/2}^2 - 1)(t_r^2/z_{\alpha/2}^2 - 1) \geqslant c; \tag{5}$$

see Appendix A for a derivation. Significance of the original study implies that $z_{\alpha/2}^2 < t_o^2$ holds; therefore $z_{\alpha/2}^2 < t_r^2$ must also hold to ensure that the left-hand side of equation (5) is positive. The required squared quantile $z_S^2 = z_{p_S/2}^2$ to obtain equality in expression (5) must therefore fulfil

$$0 \leqslant z_S^2 < \min\{t_o^2, t_r^2\} \tag{6}$$

and defines the sceptical p-value $p_S = 2\{1 - \Phi(|z_S|)\}$ via

$$(t_o^2/z_S^2 - 1)(t_r^2/z_S^2 - 1) = c. \tag{7}$$

The requirement $p_S \leqslant \alpha$ for replication success at level α now translates to $z_S^2 \geqslant z_{\alpha/2}^2$.

Equation (7) can be rewritten as

$$(c-1)z_S^4 + 2z_S^2 t_A^2 = t_A^2 t_H^2, \tag{8}$$

where $t_A^2 = (t_o^2 + t_r^2)/2$ is the arithmetical and $t_H^2 = 2/(1/t_o^2 + 1/t_r^2)$ the harmonic mean of the squared test statistics t_o^2 and t_r^2. The only solution of equation (8) that fulfils inequality (6) is

$$z_S^2 = \begin{cases} t_H^2/2 & \text{for } c = 1, \\ \dfrac{1}{c-1}\{\sqrt{[t_A^2\{t_A^2 + (c-1)t_H^2\}]} - t_A^2\} & \text{for } c \neq 1. \end{cases} \tag{9}$$

In the introductory example the original and the replication CIs have the same width, so $c = 1$ and z_S^2 is simply half the harmonic mean of $t_o^2 = 12.19$ and $t_r^2 = 3.99$, i.e. $z_S^2 = 3.00$, $|z_S| = 1.73$ and $p_S = 2\{1 - \Phi(1.73)\} = 0.083$. We can thus declare replication success at any prespecified level $\alpha \geqslant 0.083$.

3.1. Properties

Inequalities (6) imply that the sceptical p-value p_S is always larger than both the original and the replication p-values p_o and p_r. Closer inspection of equation (7) shows that z_S^2 is increasing

with increasing t_o^2 (for fixed t_r^2 and c) and also with increasing t_r^2 (for fixed t_o^2 and c). Therefore, the smaller p_o (or p_r) is, the smaller p_S (for fixed c). Furthermore, for fixed test statistics t_o and t_r (so fixed p-values p_o and p_r), the solution z_S^2 of equation (7) will decrease with increasing variance ratio

$$c = \frac{\sigma_o^2}{\sigma_r^2} = \frac{t_r^2/t_o^2}{\hat{\theta}_r^2/\hat{\theta}_o^2}.$$

Since t_r^2/t_o^2 is fixed, c increases with decreasing squared effect size ratio $\hat{\theta}_r^2/\hat{\theta}_o^2$. In other words, for the same ordinary p-values p_o and p_r, the sceptical p-value p_S increases with decreasing absolute replication effect estimate relative to the original effect estimate. This is a desired property, as replication studies with smaller effect estimates than the original estimates are considered less credible (Simonsohn, 2015).

To illustrate the dependence of p_S on the variance ratio c, consider a scenario where $p_o = p_r = 0.01$, so $t_o^2 = t_r^2$ and therefore $c = \hat{\theta}_o^2/\hat{\theta}_r^2$. First assume equal effect sizes $\hat{\theta}_r = \hat{\theta}_o$, so $c = 1$. The sceptical p-value turns out to be $p_S = 0.069$. For $\hat{\theta}_r = \hat{\theta}_o/2$ ($c = 4$) we obtain a larger value ($p_S = 0.14$) because the effect estimate $\hat{\theta}_r$ of the replication study is just half as large as the original estimate $\hat{\theta}_o$. In contrast, for $\hat{\theta}_r = 2\hat{\theta}_o$ ($c = \frac{1}{4}$) the sceptical p-value becomes smaller ($p_S = 0.035$). This asymmetry in the incorporation of the original and replication study data is natural, placing less weight on replication studies with relatively small effect estimates. This is so in the introductory example, where substantial shrinkage of the replication effect estimate leads to a relatively large sceptical p-value.

It is also interesting to study limiting values of the sceptical p-value. If we let $\sigma_o^2 \downarrow 0$ for fixed $\hat{\theta}_o \neq 0$, equation (8) reduces to the requirement $z_S^2 = t_r^2$, as shown in Appendix B. Thus, the ordinary p-value of the replication study is a special case of the sceptical p-value if the uncertainty of the original effect estimate is ignored. In contrast, ignoring the uncertainty of $\hat{\theta}_r \neq 0$ via $\sigma_r^2 \downarrow 0$ leads to $z_S^2 \uparrow z_M^2$ where

$$z_M^2 = \frac{\sqrt{\{d(d+4)\}} - d}{2} t_o^2, \tag{10}$$

with $d = \hat{\theta}_r^2/\hat{\theta}_o^2$; see Appendix B for a proof. Using the criterion $z_M^2 \geqslant z_{\alpha/2}^2$ rather than $z_S^2 \geqslant z_{\alpha/2}^2$ to assess replication success corresponds to the Matthews (2001a, b) approach that was mentioned in Section 2.2. For any value of d, z_M^2 is smaller than t_o^2 but can be larger than t_r^2. Ignoring the uncertainty of the replication effect estimate may thus lead to the declaration of replication success, even if the replication study is not conventionally significant on its own.

We may also consider the case $c \downarrow 0$ for fixed t_o^2 and t_r^2, where expression (9) increases with limit

$$z_S^2 \uparrow \min\{t_o^2, t_r^2\}, \tag{11}$$

as shown in Appendix C. Therefore $p_S \downarrow \max\{p_o, p_r\}$ for $c \downarrow 0$, which we shall use in Section 3.4.

3.2. Relationship to intrinsic credibility

The concept of intrinsic credibility has been proposed in Matthews (2018) and adapted in Held (2019) to check the credibility of 'out of the blue' findings without any prior support. In the present context this corresponds to an original study in the absence of a replication study. The idea is to evaluate the credibility of the original study if we could observe exactly the same result in the replication study.

The approach by Matthews (2018) corresponds to the case where we ignore the uncertainty of the (hypothetical) replication study and thus leads to equation (10) with $d = 1$: $z_M^2 = \{(\sqrt{5}-1)/2\}t_o^2 \approx 0.618t_o^2$. However, if we incorporate the uncertainty by using the prior-predictive approach by Box (1980), then we obtain $z_S^2 = 0.5t_o^2$ as a special case of expression (9) for $c = 1$ and $t_o^2 = t_r^2$. Now p_S reduces to the p-value for intrinsic credibility,

$$p_{IC} = 2\{1 - \Phi(t_o/\sqrt{2})\}, \tag{12}$$

as proposed in Held (2019) for the assessment of claims of new discoveries. Intrinsic credibility at level α is achieved if $p_{IC} \leqslant \alpha$, i.e. $t_o^2 \geqslant 2z_{\alpha/2}^2$, which is equivalent to $p_o \leqslant \alpha_{IC}$, where

$$\alpha_{IC} = 2\{1 - \Phi(\sqrt{2}z_{\alpha/2})\} \tag{13}$$

is the p-value threshold for intrinsic credibility. For $\alpha = 0.05$ we obtain $\alpha_{IC} = 0.0056$; for $\alpha = 0.10$ we have $\alpha_{IC} = 0.02$. These thresholds will become important in Section 4.

3.3. One-sided sceptical p-values

The procedure that was described above is designed for standard two-sided CIs and assesses replication success in a two-sided fashion, as the sign of $\hat{\theta}_r$ does not matter in the computation of the sceptical p-value. In extreme cases, it may therefore happen that a replication study is classified as successful although the signs of $\hat{\theta}_o$ and $\hat{\theta}_r$ differ. This 'replication paradox' may also occur in a Bayes factor approach; see Ly *et al.* (2018) for details.

It is therefore of interest to adapt the sceptical p-value to the one-sided setting. Without loss of generality consider the one-sided alternative $H_1 : \theta > 0$ to $H_0 : \theta = 0$ and assume that $\hat{\theta}_o > 0$. We now start with a one-sided CI for θ at level $1 - \tilde{\alpha}$ whose lower limit $\hat{\theta}_o - z_{\tilde{\alpha}}\sigma_o$ equals the lower limit L of the corresponding two-sided CI at level $1 - 2\tilde{\alpha}$. The variance τ^2 of the sufficiently sceptical prior therefore is equation (2) with $z_{\alpha/2}$ replaced by $z_{\tilde{\alpha}}$.

The obvious one-sided requirement for replication success $t_{Box} = \hat{\theta}_r/\sqrt{(\tau^2 + \sigma_r^2)} \geqslant z_{\tilde{\alpha}}$ now replaces the two-sided requirement $t_{Box}^2 \geqslant z_{\alpha/2}^2$ and ensures that the replication paradox cannot occur. If $\hat{\theta}_r \geqslant 0$, we can hence still use expression (9) to compute z_S^2 from t_o, t_r and c and the one-sided sceptical p-value turns out to be $\tilde{p}_S = 1 - \Phi(z_S)$, so half of the two-sided sceptical p-value: $\tilde{p}_S = p_S/2$. The same relationship holds between ordinary one- and two-sided p-values, of course, which implies that the one-sided sceptical p-value $\tilde{p}_S$ is always larger than the ordinary one-sided p-values $\tilde{p}_o$ and $\tilde{p}_r$ from the two studies. If the replication effect estimate is in the wrong direction, i.e. $\hat{\theta}_r < 0$, it is natural to set $\tilde{p}_S = 1 - p_S/2$.

One-sided sceptical p-values are appropriate if the study protocol of the original study is already formulated in a one-sided fashion. A *post hoc* (after the original study result is known) formulation of a one-sided alternative would require halving the original two-sided significance level α to $\tilde{\alpha} = \alpha/2$. The one-sided assessment of replication success at level $\tilde{\alpha}$ is then equivalent to the two-sided procedure at the original level α, if the signs of the original and replication effect estimates agree. This procedure ensures that the replication paradox cannot occur.

However, the one-sided p-value mapping (from $\tilde{p}_o$ and $\tilde{p}_r$ to $\tilde{p}_S$) will be different from the corresponding two-sided mapping (from p_o and p_r to p_S) because the same ordinary one- and two-sided p-values correspond to different test statistics t_o and t_r. For numerical illustration suppose that $\tilde{p}_o = \tilde{p}_r = 0.01$ (so $p_o = p_r = 0.02$ and $t_o = t_r = 2.33$) and that $\hat{\theta}_o$ and $\hat{\theta}_r$ are both positive. Then $\tilde{p}_S = 0.05$ for $c = 1$, $\tilde{p}_S = 0.09$ for $c = 4$ and $\tilde{p}_S = 0.029$ for $c = \frac{1}{4}$. These one-sided sceptical p-values are slightly smaller than the two-sided sceptical p-values for two-sided $p_o = p_r = 0.01$ (where $t_o = t_r = 2.58$), as reported in Section 3.1 ($p_S = 0.069$, 0.14, 0.035 for $c = 1$, 4, $\frac{1}{4}$ respectively). This illustrates that the one-sided assessment of replication success based on

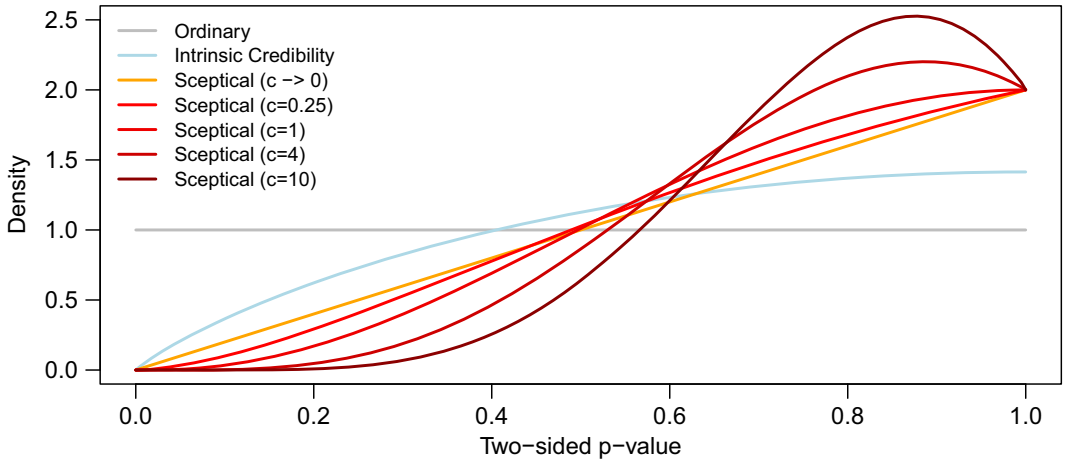

Fig. 2. Density function of the sceptical p-value for various values of the variance ratio c under the assumption of no effect: the density for the limiting case $c \longrightarrow 0$ as well as the density of the p-value for intrinsic credibility and the ordinary p-value are also shown

one-sided ordinary p-values is slightly less stringent than the two-sided assessment based on two-sided ordinary p-values, if the original and replication effect estimates have the same sign.

3.4. The distribution under the null

It is interesting to compare the distributions of p_o (or p_r), p_{IC} and p_S under the assumption of no effect, where the ordinary p-value is uniformly distributed. We can easily derive the density of p_{IC} with a change-of-variables by using equation (12): $f(p_{IC}) = 2\sqrt{\pi}\varphi\{t(p_{IC})\}$; here $\varphi(\cdot)$ is the standard normal density function and $t(p_{IC}) = \Phi^{-1}(1 - p_{IC}/2)$.

The distribution of p_S can be studied via stochastic simulation. Density estimates are displayed in Fig. 2 for various values of the variance ratio c based on 5×10^6 samples each. We can see that the risk of small 'false positive' sceptical p-values is drastically reduced, compared with ordinary p-values based on one study only. Note that the variance is usually inversely proportional to the sample size of each study, i.e. $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$ for some unit variance κ^2, say. Then $c = n_r/n_o$, so the variance factor c is increasing with increasing sample size n_r of the replication study. The distribution of p_S in Fig. 2 is shifted to the right with increasing c, so an increasing sample size of the replication study reduces the risk of a false claim of replication success.

From expression (11) we know that for $c \downarrow 0$ we have $p_S \downarrow \max\{p_o, p_r\}$, which follows a triangular Be(2, 1) distribution if p_o and p_r are independently uniform. The corresponding density function is shown in Fig. 2, as well as the density function of p_o and p_{IC}. The triangular distribution gives the upper bound α^2 for the tail probability $\Pr(p_S \leqslant \alpha \mid H_0)$ for sufficiently small α and any value of the variance ratio c. For example, for $\alpha = 0.05$ we obtain $\Pr(p_S \leqslant 0.05 \mid H_0) \leqslant 0.0025$ for any c. This is to be compared with $\Pr(p_o \leqslant 0.05 \mid H_0) = 0.05$ and $\Pr(p_{IC} \leqslant 0.05 \mid H_0) = 0.0056$. However, α^2 is not a particularly sharp bound. If, for example, the replication sample size equals the original sample size ($c = 1$), then $\Pr(p_S \leqslant 0.05 \mid H_0) \approx 0.0001$ is much smaller than 0.0025.

4. Power and sample size calculations

Replication success is not only a function of the two p-values from the original and replication study, but also of sample size, which enters in the variance ratio c. The computation of the power

or the required replication sample size to achieve replication success is hence more challenging than in standard sample size calculations. A larger sample size will be required since replication success (which is defined as $p_S \leqslant \alpha$) implies significance of the replication study ($p_r \leqslant \alpha$). Furthermore, the sample size required will depend on the p-value p_o from the original study.

The Bayesian assessment of sample size uses a design prior (O'Hagan and Stevens, 2001; O'Hagan *et al.*, 2005) to express prior beliefs about the true effect size. To power a study for replication success, the results from the original study will thus enter in two ways: as design prior for the effect size and in the subsequent assessment of replication success. For the former I shall distinguish two cases: a normal prior with mean $\hat{\theta}_o$ and variance σ_o^2 and a point prior at $\hat{\theta}_o$. The normal prior incorporates the uncertainty of $\hat{\theta}_o$ whereas the point prior does not, by analogy with the concepts of predictive and conditional power in clinical trials (Spiegelhalter and Freedman, 1986; Spiegelhalter *et al.*, 1986).

Suppose that n_o is the size of the original study sample and n_r the sample size of the replication study, so $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$, where κ^2 is the unit variance from one observation. Then $c = n_r/n_o$, which would also hold in a balanced two-sample design with respective sample sizes n_o and n_r per group. Under an initial uniform prior for θ, the sampling distribution $\hat{\theta}_o \sim N(\theta, \sigma_o^2)$ of the original study now serves as prior distribution $\theta \mid \hat{\theta}_o \sim N(\hat{\theta}_o, \sigma_o^2)$ with prior-predictive distribution

$$\hat{\theta}_r \mid \hat{\theta}_o \sim N\left\{\hat{\theta}_o, \sigma_o^2 + \sigma_r^2 = \kappa^2\left(\frac{1}{n_o} + \frac{1}{n_r}\right)\right\} \tag{14}$$

for the observed effect $\hat{\theta}_r$ in the replication study. Then t_r^2 follows a scaled non-central χ^2-distribution with 1 degree of freedom, scaling factor $1 + c$ and non-centrality parameter $t_o^2/(1 + 1/c)$, as shown in Appendix D. For the alternative point prior at $\theta = \hat{\theta}_o$, t_r^2 follows a non-central χ^2-distribution with 1 degree of freedom and non-centrality parameter $\lambda = ct_o^2$.

To compute the power for replication success, the relative sample size $c = n_r/n_o$ in equation (9) is fixed. Then z_S^2 and the sceptical p-value p_S are monotone functions of t_r^2 and we can compute the power for replication success at any level α. We can also calculate the required relative sample size $c = n_r/n_o$ at some predefined power for replication success. Both tasks require application of numerical root finding algorithms. Computational details are omitted here.

4.1. Power calculations

Fig. 3 compares the power for significance with the power for replication success for a replication study with sample size equal to the original study ($c = 1$) at level $\alpha = 5\%$ as a function of the two-sided p-value p_o of the original study. Power calculations for significance aim to detect the effect estimate $\hat{\theta}_o$ from the original study with a standard two-sided significance test. Not accounting for the associated uncertainty corresponds to the concept of conditional power, whereas predictive power calculations are based on a normal design prior with mean $\hat{\theta}_o$ and variance σ_o^2 (Spiegelhalter *et al.* (2004), equation (6.4)). The results are in accordance with those reported in Goodman (1992). In particular, for $p_o = 0.05$ the power is 50% both for conditional and for predictive power with conditional power increasing faster than predictive power for smaller p-values p_o.

Conditional and predictive power for replication success is also shown in Fig. 3 for two-sided $\alpha = 5\%$ and one-sided $\tilde{\alpha} = 5\%$. As expected, the power for replication success is lower than for significance and drops quickly to 0 for values of p_o that are close to 0.05. Remarkably, in the two-sided case, a conditional and predictive power of 50% is attained at $p_o = 0.0056$. This is the threshold (13) for intrinsic credibility at level $\alpha = 5\%$, as described in Section 3.2. In the

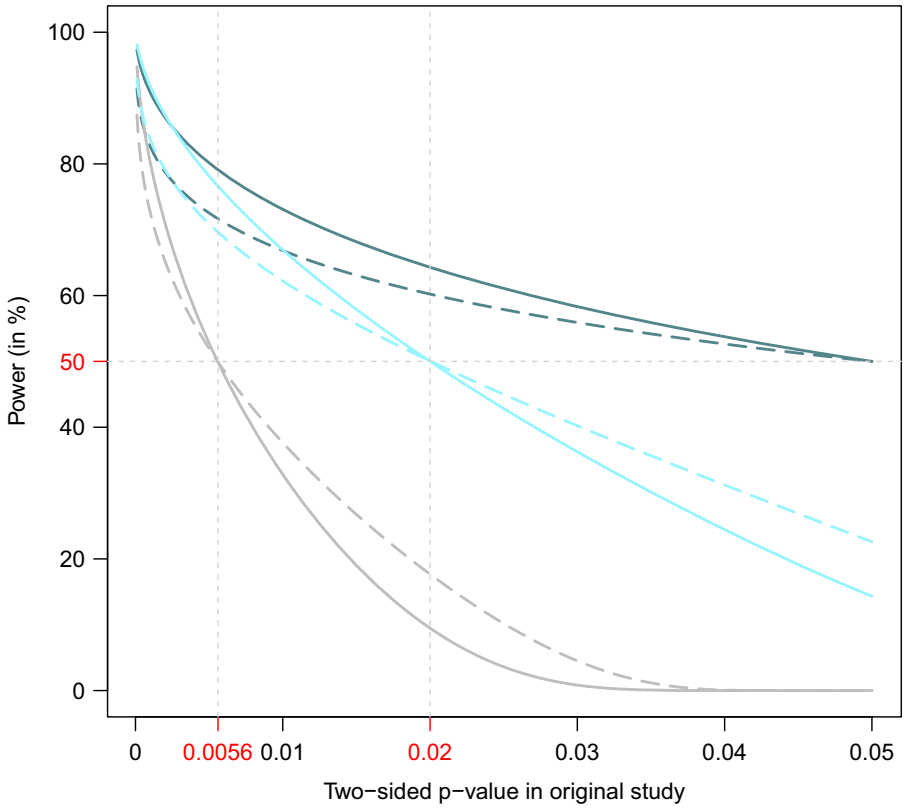

Fig. 3. Power calculations for a replication study with sample size equal to the original study ($c = 1$): shown is conditional (—, —, —) and predictive (— —, — —, — —) power for significance (two-sided) (—, — —) and for replication success (—, — —, one sided; —, — —, two sided) at level $\alpha = 5\%$ as a function of the two-sided p-value of the original study; the one- and two-sided thresholds for intrinsic credibility are marked in red on the x-axis

one-sided case a power of 50% is obtained at $p_o = 0.02$: the threshold for intrinsic credibility at two-sided level $\alpha = 10\%$. Therefore, only intrinsically credible results (based on the threshold (13)) ensure that the power for success of an identically designed replication study exceeds 50%. This intriguing feature highlights the difficulty of achieving replication success if the evidence from the original study is only suggestive and provides a new argument for more stringent p-value thresholds for claims of new discoveries (Johnson, 2013; Benjamin *et al.*, 2018; Ioannidis, 2018; Held, 2019).

This surprising result can be explained as follows: if the non-centrality parameter λ of a non-central $\chi^2(1)$ distribution is reasonably large, say $\lambda > 4$, then the median is approximately equal to λ. Under the point prior and for $c = 1$, the non-centrality parameter of the distribution of t_r^2 is $\lambda = t_o^2$, so $\text{Med}(t_r^2) \approx t_o^2$. The sceptical p-value $p_S = 2\{1 - \Phi(z_S)\}$ is then defined through $z_S^2 = t_H^2/2 = (1/t_o^2 + 1/t_r^2)^{-1}$, so the median of z_S^2 is approximately $t_o^2/2$. Replication success is thus achieved with 50% probability for $z_{\alpha/2}^2 = z_S^2 \approx t_o^2/2$, i.e. $t_o^2 \approx 2z_{\alpha/2}^2$. This corresponds to the intrinsic credibility threshold (13) for the ordinary p-value p_o from the original study.

We obtain essentially the same result under the normal prior, where now $\lambda = t_o^2/2$, which combined with a scaling factor of 2 also leads to $\text{Med}(t_r^2) \approx t_o^2$ for sufficiently large t_o^2. The rest of the argument is as above, but note that this approximation is slightly less precise, because the

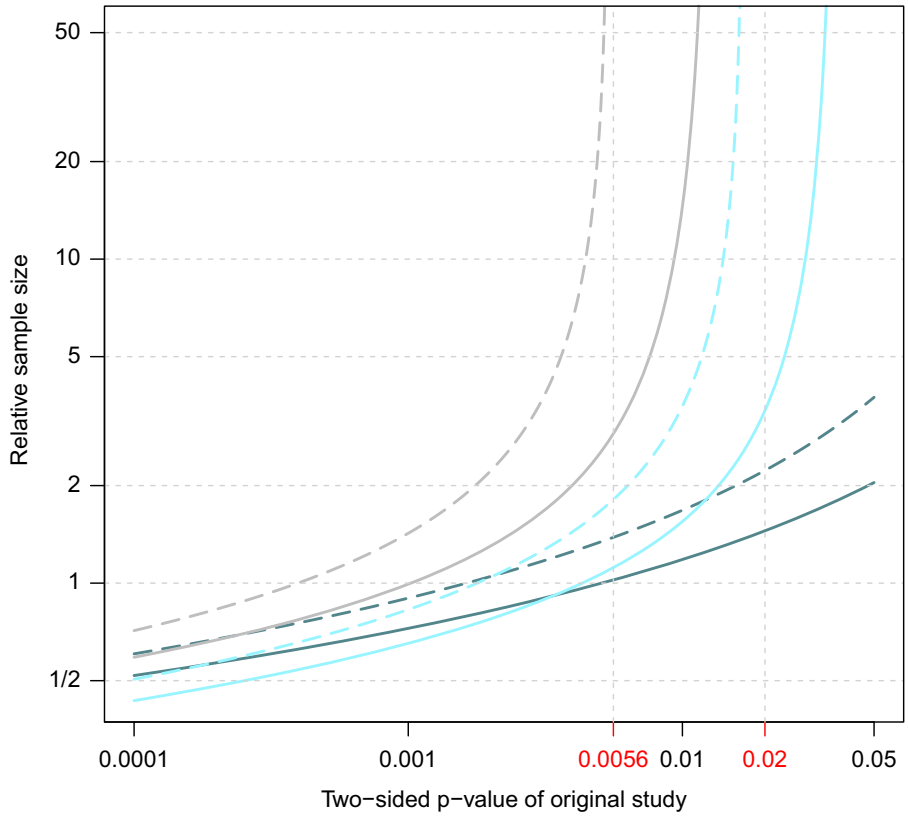

Fig. 4. Relative sample size $c = n_r/n_o$ to achieve significance (——, two sided) and replication success (——, one sided; ——, two sided) with 80% conditional (——, ——, ——) or predictive (— —, — —, — —) power for $\alpha = 0.05$ as a function of the original two-sided p-value p_o: the one- and two-sided thresholds for intrinsic credibility are marked in red on the x-axis

non-centrality parameter is half as large as under the point prior. The approximation is, however, still very good: for $\alpha = 5\%$, the exact power for replication success at $p_o = 2\{1 - \Phi(\sqrt{2}z_{0.025})\}$ is 50.00001% for the point prior and 50.00459% for the normal prior.

4.2. Sample size calculations

Fig. 4 compares various strategies to determine the replication sample size for two-sided $\alpha = 5\%$ and one-sided $\tilde{\alpha} = 5\%$ and original two-sided p-values p_o between 0.0001 and 0.05. The power to achieve significance and replication success is fixed at 80%.

Standard (two-sided) sample size calculations based on conditional power give relative sample sizes between 0.52 and 2, depending on the p-value p_o of the original study. Incorporating the uncertainty from the original study based on predictive power gives relative sample sizes between 0.61 and 3.7.

The required relative sample size for two-sided replication success is larger than that for significance alone and depends more drastically on the p-value p_o of the original study. First consider the case of conditional power. If p_o is smaller than 0.001, the relative sample size c required is smaller than 1, so the replication sample size n_r does not need to be larger than the original sample size n_o. However, the sample size required explodes for larger p-values

with an asymptote around $p_o = 0.012$. This highlights the difficulty of achieving 80% power for replication success with original p-values between 0.01 and 0.05. Even larger sample sizes are required based on predictive rather than conditional power with an asymptote around $p_o = 0.005$.

The curves shift a little to the right when we assess replication success in a one-sided fashion, pushing the asymptotes towards $p_o = 0.035$ for conditional and $p_o = 0.017$ for predictive power. The predictive power asymptotes are remarkably close to the corresponding thresholds 0.0056 and 0.02 for intrinsic credibility, which are also shown in Fig. 4. Of course, the asymptotes would change for power values that are different from 80%.

5. Replication success in psychological science

I now reanalyse data from the Open Science Collaboration (2015) project, which is a multi-year endeavour that replicated 100 scientific studies that were selected from three prominent psychology journals. Effect sizes have been transformed to correlation coefficients $\hat{\rho}$ where application of Fisher's z-transformation $\hat{\theta} = \tanh^{-1}(\hat{\rho})$ justifies a normal assumption with the

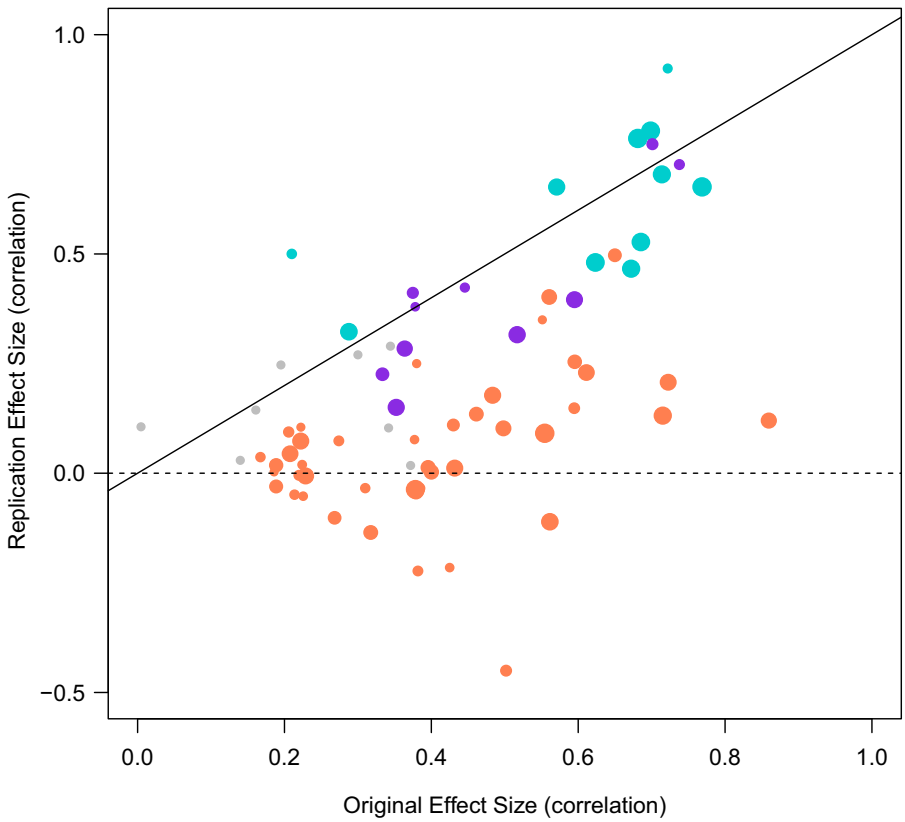

Fig. 5. Application to the Open Science Collaboration (2015) data: the circles represent the effect estimates (correlations) of the original and replication studies; the size of the circles represents the predictive power for replication success at the two-sided 5% level (small, 20%; medium, 50%; large, 80%); replication success and significance are also assessed at the two-sided 5% level and indicated by the colour of the circles (●, replication success ($n = 11$); ●, both studies significant, but no replication success ($n = 10$); ●, only orginal study significant ($n = 44$); ●, original study not significant ($n = 8$))

Table 1. Results for the 24 most successful replication studies with $p_S \leqslant 0.15$, as listed in the last column†

Study	*Results for original study*			*Results for replication study*			*Results for replication success*	
	n_o	$\hat{\rho}_o$	p_o	n_r	$\hat{\rho}_r$	p_r	*Power (%)*	p_S
1	126	0.68	<0.0001	177	0.76	<0.0001	>99.9	<0.0001
2	78	0.77	<0.0001	38	0.65	<0.0001	>99.9	<0.0001
3	30	0.70	<0.0001	31	0.78	<0.0001	95.3	0.0005
4	174	0.29	0.0001	141	0.32	<0.0001	82.6	0.005
5	32	0.57	0.0005	32	0.65	<0.0001	78.7	0.007
6	22	0.71	<0.0001	22	0.68	0.0003	87.6	0.008
7	38	0.62	<0.0001	39	0.48	0.002	93.7	0.011
8	30	0.69	<0.0001	27	0.53	0.004	92.1	0.015
9	117	0.21	0.023	236	0.50	<0.0001	14.0	0.033
10	23	0.67	0.0003	31	0.47	0.007	88.4	0.038
11	9	0.72	0.026	18	0.92	<0.0001	9.3	0.048
12	154	0.36	<0.0001	50	0.28	0.045	69.7	0.052
13	40	0.37	0.017	95	0.41	<0.0001	29.0	0.06
14	11	0.70	0.014	11	0.75	0.006	28.9	0.067
15	25	0.59	0.001	33	0.40	0.022	76.7	0.072
16	41	0.52	0.0004	41	0.32	0.044	79.6	0.08
17	9	0.74	0.021	16	0.70	0.002	18.9	0.091
18	33	0.56	0.0005	21	0.40	0.071	64.5	0.096
19	33	0.38	0.029	72	0.38	0.0009	4.3	0.10
20	25	0.45	0.025	39	0.42	0.007	10.9	0.10
21	57	0.33	0.011	118	0.23	0.014	44.5	0.11
22	96	0.20	0.057	243	0.25	<0.0001	0.0	0.12
23	16	0.65	0.005	13	0.50	0.085	45.9	0.13
24	69	0.35	0.003	178	0.15	0.045	78.0	0.14

†The penultimate column gives the predictive power for replication success at level $\alpha = 5\%$.

standard error being a function only of the nominal study sample size n: $\text{se}(\hat{\theta}) = 1/\sqrt{(n-3)}$, so the effective sample size is $n-3$ with variance ratio $c = (n_r - 3)/(n_o - 3)$. Effective sample sizes are available for 73 studies: the so-called meta-analytic subset (Johnson *et al.*, 2016). Two-sided p-values p_o and p_r have been calculated assuming normality of the corresponding test statistics $t_o = \hat{\theta}_o/\text{se}(\hat{\theta}_o)$ and $t_r = \hat{\theta}_r/\text{se}(\hat{\theta}_r)$ respectively. I have not included the remaining 27 studies where a normal approximation is questionable since the standard errors of $\hat{\theta}_o$ and $\hat{\theta}_r$ are not available.

Fig. 5 displays the replication *versus* the original correlation estimates. Eight of the 73 original studies are not significant at the standard $\alpha = 5\%$ level, three of them with p-values between 0.05 and 0.06. There have been 21 significant replication studies following from the 65 significant original studies. The sceptical p-value enables us to rank the studies by the degree of replication success. Table 1 lists the 24 most successful replication studies with $p_S \leqslant 0.15$ of which the top 11 have been successful at the two-sided 5% level ($p_S \leqslant 0.05$). The remaining 13 studies in Table 1 (with $p_S > 0.05$) show some interesting features. For example, study 18 has a non-significant replication result but still leads to replication success at the 10% level. Conversely, there are several studies with both $p_o \leqslant 0.05$ and $p_r \leqslant 0.05$ but $p_S > 0.10$. This illustrates, once again, that the sceptical p-value not only takes significance of the original and replication study into account, but also effect and sample sizes, both entering in the variance ratio c.

Table 1 also gives the predictive power for replication success at the two-sided 5% level: a function of p_o and c only. Studies 11 and 24 are particularly interesting, both having significant original and replication results. Replication study 11 has a very low predictive power of 9.3% but still leads to $p_S = 0.048$ because the estimated correlation in the replication study is even larger than in the original study (0.92 *versus* 0.72) and the replication p-value is very small ($p_r < 0.0001$). In contrast, replication study 24 has a reasonably large power of 78.0% but leads to $p_S = 0.14$, because there is substantial shrinkage of the effect estimate (0.15 *versus* 0.35) combined with borderline significance only ($p_r = 0.045$). The two studies can be easily identified in Fig. 5.

6. Discussion

Science would proceed more efficiently if statistical approaches to inference are better aligned with scientific needs and practice (Goodman, 2016). The traditional dichotomy between 'Bayesians' and 'frequentists' may not always be useful to achieve this. The proposed methodology in this paper represents a Bayes–non-Bayes compromise (Good, 1992) for extracting more insight from replication studies based on standard metrics (effect estimates, confidence intervals and p-values). Instead of synthesizing original and replication study results through a meta-analysis, the original study result is challenged with the sufficiently sceptical prior. Replication success is then defined as conflict between the sufficiently sceptical prior and the replication effect estimate. Whereas the ordinary p-value quantifies the conflict between the point null hypothesis and the replication data, the sceptical p-value quantifies the conflict between the sufficiently sceptical prior and the replication data. It extends the ordinary p-value of the replication study by taking into account effect and sample sizes of the two studies.

Just as the ordinary p-value is an indirect measure of the evidence against the null hypothesis, the sceptical p-value is an indirect measure of the degree of replication success. Specifically, a large sceptical p-value can occur if the replication sample size was too small, even if the original and replication effect sizes are approximately equal, and should not be taken as evidence for no effect (Altman and Bland, 1995). This is not the only reason why it would be interesting to compare the sceptical p-value with direct 'forward Bayes' approaches, such as the replication Bayes factor (Verhagen and Wagenmakers, 2014; Ly *et al.*, 2018), which quantifies the change in evidence that is brought about by observing the results from the replication study, given that the evidence from the original study is already available.

Significance of both the original and the replication study is a necessary but not sufficient requirement for replication success. The framework proposed thus extends the 'two-pivotal-studies paradigm' requiring two significant findings from two independent confirmatory trials for regulatory drug approval; see Kennedy-Shaffer (2017) for a recent review. However, the difficulty of achieving replication success if the evidence from the original study is only suggestive underlines the need for more stringent p-value thresholds for claims of new discoveries. The threshold for intrinsic credibility (13) is a natural choice for this task.

It would be interesting to extend the approach to a setting where several replication studies are available. For example, a summary estimate based on a meta-analysis of all available replication studies may be used to assess replication success. If results from replication studies become sequentially available, an alternative approach is first to combine original and replication effect estimates into a summary measure. Some downweighting of the original study result will in general be required depending on the degree of conflict between the original and replication study result, e.g. with adaptive power priors (Gravestock and Held, 2017). The summary measure could then be used as a new 'original' effect estimate to assess the success of a second replication study. This would open up new ways to challenge existing knowledge iteratively through a

series of replication studies and would provide an interesting alternative to traditional evidence synthesis methods.

The reverse Bayes approach proposed assumes a simple mathematical framework, where likelihood, prior and posterior are all assumed to be normal. It will be of interest to extend this framework to other settings, e.g. to the binomial or t-distribution.

7. Data and software availability

Data that were analysed in this paper were originally from Open Science Collaboration (2015) and have been downloaded from `https://osf.io/fgjvw/`. Software to compute the sceptical p-value and the power or required sample size to achieve replication success are available in the R package `ReplicationSuccess` which is available from `R-Forge`; use the R command `install.packages("ReplicationSuccess",repos="http://R-Forge.R-project.org")`.

Acknowledgements

I am grateful to Robert Matthews, Ken Rice, Uri Simonsohn and the members of the University of Zurich Department of Biostatistics for helpful discussions and suggestions. I also acknowledge helpful comments by referees and a reviewer on a related grant proposal of mine.

Support by the Swiss National Science Foundation (project 189295) is gratefully acknowledged.

Appendix A: Proof of equation (5)

We have

$$t^2_{\text{Box}} = \frac{\hat{\theta}^2_{\text{r}}}{\tau^2+\sigma^2_{\text{r}}} = \frac{\hat{\theta}^2_{\text{r}}}{\sigma^2_{\text{r}}}\left(\frac{c}{t^2_{\text{o}}/z^2_{\alpha/2}-1}+1\right)^{-1} = t^2_{\text{r}}\frac{t^2_{\text{o}}/z^2_{\alpha/2}-1}{c+t^2_{\text{o}}/z^2_{\alpha/2}-1},$$

so the requirement $t^2_{\text{Box}} \geqslant z^2_{\alpha/2}$ for replication success at level α is equivalent to

$$\frac{t^2_{\text{r}}}{z^2_{\alpha/2}}\left(\frac{t^2_{\text{o}}}{z^2_{\alpha/2}}-1\right) \geqslant c + t^2_{\text{o}}/z^2_{\alpha/2} - 1.$$

Subtracting $t^2_{\text{o}}/z^2_{\alpha/2}-1$ on both sides leads to equation (5).

Appendix B: The limiting cases $\sigma^2_{\text{o}} \downarrow 0$ and $\sigma^2_{\text{r}} \downarrow 0$

Equation (8) can be rewritten as

$$\frac{c-1}{t^2_{\text{A}}} z^4_{\text{S}} + 2z^2_{\text{S}} = t^2_{\text{H}}, \tag{15}$$

where

$$\frac{c-1}{t^2_{\text{A}}} = \frac{\sigma^2_{\text{o}}-\sigma^2_{\text{r}}}{\sigma^2_{\text{r}}}\frac{2}{t^2_{\text{o}}+t^2_{\text{r}}} = \frac{2\sigma^2_{\text{o}}(\sigma^2_{\text{o}}-\sigma^2_{\text{r}})}{\hat{\theta}^2_{\text{o}}\sigma^2_{\text{r}}+\hat{\theta}^2_{\text{r}}\sigma^2_{\text{o}}}.$$

For $\sigma^2_{\text{o}} \downarrow 0$ we thus have $(c-1)/t^2_{\text{A}} \to 0$ and $t^2_{\text{H}} \to 2t^2_{\text{r}}$ so equation (15) reduces to $2z^2_{\text{S}} = 2t^2_{\text{r}}$ and hence $z^2_{\text{S}} = t^2_{\text{r}}$. For $\sigma^2_{\text{r}} \downarrow 0$ we have $(c-1)/t^2_{\text{A}} \to 2\sigma^2_{\text{o}}/\hat{\theta}^2_{\text{r}}$ and $t^2_{\text{H}} \to 2t^2_{\text{o}}$ so equation (15) reduces to $(\sigma^2_{\text{o}}/\hat{\theta}^2_{\text{r}})z^4_{\text{S}} + z^2_{\text{S}} = t^2_{\text{o}}$. The solution of this equation is

$$
\begin{aligned}
z_S^2 &= \frac{\hat\theta_r^2}{2\sigma_o^2}\{\sqrt{(1+4\sigma_o^2 t_o^2/\hat\theta_r^2)} - 1\} \\
&= \frac{\hat\theta_r^2}{2\sigma_o^2}\{\sqrt{(1+4/d)} - 1\} \\
&= \frac{\hat\theta_o^2}{2\sigma_o^2}\{\sqrt{(d^2+4d)} - d\} \\
&= \frac{t_o^2}{2}[\sqrt{\{d(d+4)\}} - d],
\end{aligned}
$$

which is equation (10) with $d = \hat\theta_r^2/\hat\theta_o^2$.

Appendix C: Proof of result (11)

Equation (9) reduces for $c \downarrow 0$ to

$$
z_S^2 = t_A^2 - \sqrt{\{t_A^2(t_A^2 - t_H^2)\}} = t_A^2 - \sqrt{\left\{\frac{t_A^2}{2}\frac{(t_o^2 - t_r^2)^2}{t_o^2 + t_r^2}\right\}} = t_A^2 - \frac{|t_o^2 - t_r^2|}{2} = \min\{t_o^2, t_r^2\}.
$$

The derivative of equation (9) with respect to c is (for $c \neq 1$)

$$
\begin{aligned}
\frac{\mathrm{d}z_S^2}{\mathrm{d}c} &= -\frac{1}{c-1}\left\{z_S^2 - \frac{1}{2}\frac{t_A^2 t_H^2}{(c-1)z_S^2 + t_A^2}\right\} \\
&= -\frac{z_S^2}{c-1}\left\{1 - \frac{1}{2}\frac{(c-1)z_S^2 + 2t_A^2}{(c-1)z_S^2 + t_A^2}\right\} \\
&= -\frac{1}{2}\frac{z_S^4}{(c-1)z_S^2 + t_A^2} \qquad (16)
\end{aligned}
$$

where the middle line follows from equation (8) and the last line also holds for $c = 1$. It is easy to see from equation (9) that $(c-1)z_S^2 + t_A^2 > 0$ for all c, and therefore equation (16) is negative for all c.

Appendix D: Proof of results in Section 4

For notational simplicity I omit the conditioning on $\hat\theta_o$ in what follows. Equation (14) implies a distribution on $t_r = \hat\theta_r/\sigma_r = \sqrt{n_r}\hat\theta_r/\kappa$,

$$
t_r \sim N\left(\sqrt{n_r}\frac{\hat\theta_o}{\kappa}, \frac{n_o + n_r}{n_o}\right),
$$

so $t_r = \sqrt{\{(n_o + n_r)/n_o\}}\tilde t_r$ where

$$
\tilde t_r \sim N\left\{\sqrt{\left(\frac{n_o n_r}{n_o + n_r}\right)}\frac{\hat\theta_o}{\kappa}, 1\right\}.
$$

Therefore $t_r^2 = \{(n_o + n_r)/n_o\}\tilde t_r^2$ follows a scaled non-central χ^2-distribution with 1 degree of freedom, scaling factor $(n_o + n_r)/n_o = 1 + c$ and non-centrality parameter $\lambda = \{(n_o n_r)/(n_o + n_r)\}\hat\theta_o^2/\kappa^2 = t_o^2/(1 + 1/c)$.

Things simplify somewhat for a point prior $\theta = \hat\theta_o$ at the estimate from the original study. Then $\hat\theta_r \mid \hat\theta_o \sim N(\hat\theta_o, \kappa^2/n_r)$ so $t_r \sim N(\sqrt{n_r}\hat\theta_o/\kappa, 1)$. Now t_r^2 follows a non-central χ^2-distribution with 1 degree of freedom and non-centrality parameter $\lambda = n_r\hat\theta_o^2/\kappa^2 = ct_o^2$.

References

van Aert, R. C. M. and van Assen, M. A. L. M. (2017) Bayesian evaluation of effect size after replicating an original study. *PLOS One*, **12**, no. 4, 1–23.

Altman, D. G. and Bland, J. M. (1995) Absence of evidence is not evidence of absence. *Br. Med. J.*, **311**, 485.

Anderson, S. F. and Maxwell, S. E. (2017) Addressing the "replication crisis": using original studies to design replication studies with appropriate statistical power. *Multiv. Behav. Res.*, **52**, 305–324.

Bayarri, M. J. and Mayoral, M. (2002) Bayesian design of "successful" replications. *Am. Statistn*, **56**, 207–214.

Begley, C. G. and Ioannidis, J. P. (2015) Reproducibility in science. *Circ. Res.*, **116**, 116–126.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Holland Jones, J., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. (2018) Redefine statistical significance. *Nat. Hum. Behav.*, **2**, 6–10.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neursci.*, **14**, 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016) Evaluating replicability of laboratory experiments in economics. *Science*, **351**, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and Wu, H. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.*, **2**, 637–644.

Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values. *R. Soc. Open Sci.*, **4**, no. 12.

Colquhoun, D. (2019) The false positive risk: a proposal concerning what to do about *p*-values. *Am. Statistn*, **73**, suppl. 1, 192–201.

Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. *Baysn Anal.*, **1**, 893–914.

Fletcher, A., Spiegelhalter, D., Staessen, J., Thijs, L. and Bulpitt, C. (1993) Implications for trials in progress of publication of positive results. *Lancet*, **342**, 653–657.

Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.

Good, I. J. (1983) *Good Thinking: the Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Good, I. J. (1992) The Bayes/non-Bayes compromise: a brief review. *J. Am. Statist. Ass.*, **87**, 597–606.

Goodman, S. N. (1992) A comment on replication, p-values and evidence. *Statist. Med.*, **11**, 875–879.

Goodman, S. N. (2016) Aligning statistical and scientific reasoning. *Science*, **352**, 1180–1181.

Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016) What does research reproducibility mean? *Sci. Transnl Med.*, **8**, article 341ps12.

Gravestock, I. and Held, L. (2017) Adaptive power priors with empirical Bayes for clinical trials. *Pharm. Statist.*, **16**, 349–360.

Greenland, S. (2006) Bayesian perspectives for epidemiological research: I, Foundations and basic methods. *Int. J. Epidem.*, **35**, 765–775.

Greenland, S. (2011) Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev. Med.*, **53**, 225–228.

Held, L. (2013) Reverse-Bayes analysis of two common misinterpretations of significance tests. *Clin. Trials*, **10**, 236–242.

Held, L. (2019) The assessment of intrinsic credibility and a new argument for $p < 0.005$. *R. Soc. Open Sci.*, **6**, no. 3.

Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLOS Med.*, **2**, no. 8, article e124.

Ioannidis, J. P. A. (2018) The proposal to lower *p*-value thresholds to .005. *J. Am. Med. Ass.*, **319**, 1429–1430.

Johnson, V. E. (2013) Revised standards for statistical evidence. *Proc. Natn. Acad. Sci. USA*, **110**, 19313–19317.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. (2016) On the reproducibility of psychological science. *J. Am. Statist. Ass.*, **112**, 1–10.

Kennedy-Shaffer, L. (2017) When the alpha is the omega: *p*-values, "substantial evidence," and the 0.05 standard at FDA. *Food Drug Law J.*, **72**, 595–635.

Ly, A., Etz, A., Marsman, M. and Wagenmakers, E.-J. (2018) Replication Bayes factors from evidence updating. *Behav. Res. Meth.*, to be published.
Matthews, R. A. J. (2001a) Methods for assessing the credibility of clinical trial outcomes. *Drug Inform. J.*, **35**, 1469–1478.
Matthews, R. A. J. (2001b) Why *should* clinicians care about Bayesian methods (with discussion)? *J. Statist. Planng Inf.*, **94**, 43–71.
Matthews, R. A. J. (2018) Beyond 'significance': principles and practice of the analysis of credibility. *R. Soc. Open Sci.*, **5**, no. 1.
O'Hagan, A. and Stevens, J. W. (2001) Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med. Decsn Makng*, **21**, 219–230.
O'Hagan, A., Stevens, J. W. and Campbell, M. J. (2005) Assurance in clinical trial design. *Pharm. Statist.*, **4**, 187–201.
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, **349**, article aac4716.
Patil, P., Peng, R. D. and Leek, J. T. (2016) What should researchers expect when they replicate studies?: A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.*, **11**, 539–544.
Schönbrodt, F. D. and Wagenmakers, E.-J. (2018) Bayes factor design analysis: planning for compelling evidence. *Psychnom. Bull. Rev.*, **25**, 128–142.
Simonsohn, U. (2015) Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.*, **26**, 559–569.
Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. New York: Wiley.
Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.
Spiegelhalter, D. J., Freedman, L. S. and Blackburn, P. R. (1986) Monitoring clinical trials: conditional or predictive power? *Contr. Clin. Trials*, **7**, 8–17.
Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1994) Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc.* A, **157**, 357–416.
Verhagen, J. and Wagenmakers, E.-J. (2014) Bayesian tests to quantify the result of a replication attempt. *J. Exptl Psychol.*, **143**, 1457–1475.